CoFiH: A heuristic for concept discovery in computer-assisted conceptual analysis

Louis Chartrand[1], Jean-Guy Meunier[1], Davide Pulizzotto[1], José López González[2], Jean-François Chartier[1], Ngoc Tan Le, Francis Lareau, Julian Trujillo Amaya[3]

[1]Université du Québec à Montréal – Canada

[2]Université du Québec à Trois-Rivière – Canada

[2]Universidad del Valle – Colombia

Abstract

While conceptual analysis can be facilitated by computer assistance, the absence of proper models for concepts in text has curtailed the development of such tools. The most common heuristic, which consists in identifying keywords as canonical expression of a concept, poses problems of ambiguity and fails to retrieve most of the relevant textual data. In this paper, we present CoFiH, an algorithm that exploits topics in order to retrieve segments relevant to a given concept. It is then applied to C.S. Peirce's *Collected Papers* to facilitate the analysis of Peirce's concept of LAW. Compared to the baseline, CoFiH produces better recall and enables a meaningful analysis along several topics.

Key words : text mining, concept mining, conceptual analysis, computer-assisted reading, computer-assisted conceptual analysis

1. Introduction

Conceptual analysis in philosophy can refer either, in a technical sense, to the discovery of *a priori* knowledge in the concepts we share (Jackson, 1998; Laurence & Margolis, 2003) or, in a broader sense, to the philosophical methods we use to uncover the meaning and the use of a concept in order to clarify it or to improve it (Haslanger, 2012, Chapter 13). Given philosophy's focus on conceptual clarity, the latter has been ubiquitous in practice.

While a concept can also be studied through thought experiments and intuitions, much of conceptual analysis involves reading and text analysis. Indeed, it is often through texts that we can best understand how a concept has emerged and evolved into the concept we now employ, or how it is employed differently in different contexts and when used by different authors. In other words, while armchair philosophy helps one give a better account of her own concepts, contact with texts provides a necessary perspective.

However, while textual analysis has profited from various methods of computer assistance, such as lexicometry (Lebart & Salem, 1994) or text mining (Fayyad, Piatetsky-Shapiro, & Smyth, 1996; J.-F. Chartier & Meunier, 2011), philosophical conceptual analysis has remained almost untouched by these developments. As a result, few such methods have been developed, with few

exceptions which have attracted little attention from philosophers (J. G. Meunier, Forest, & Biskri, 2005; J. F. Chartier, Meunier, Danis, & Jendoubi, 2008; Danis, Meunier, Chartier, Alrahabi, & Desclés, 2010; Estève, 2008; Sainte-Marie, Meunier, Payette, & Chartier, 2011).

One important obstacle to the import of those methods in philosophy lies in the lack of proper concept models for conceptual analysis. Keywords approaches to identifying concept run into ambiguity problems, like polysemy and synonymy. Latent concept approaches such as latent semantic analysis (LSA: Deerwester, Dumais, Landauer, Furnas, & Harshman, 1990) and latent dirichlet allocation (LDA: Blei, Ng, & Jordan, 2003), which were devised in large part to address those problems, confuse the various types of semantic relationships. However, they produce a surface representation of lexical regularities, which fails to capture the intricate semantic links that a conceptual analysis cannot overlook. In general, while in conceptual analysis a philosopher ought to study all the textual contexts relevant to a concept, which often occurs in a variety of contexts, existing approaches tend to focus on a subset of the text where it is expressed, that is, a set of lexically similar textual segments where one or a few terms prototypically associated with the concept are most present.

The problem we address in this article is that of retrieving textual segments which are relevant to conceptual analysis. Our general hypothesis is that it can be achieved by a method that models concept as expressing themselves through *topics*. This enables modeled concepts to appear in different lexical contexts in different forms, and thus expand their reach. Our specific hypothesis is that, using this insight, we can produce, as a proof of concept, a heuristic that retrieves segments which are relevant to a concept of interest in such a fashion as it can facilitate conceptual analysis of this concept.

As previous approaches fail to model the kind of concept we need (section §2), our approach relies on particular notions of concept and topic which will be described both intuitively and in terms of their effect on lexicon in section §3, while the algorithm itself, CoFiH, is described in section §4. Application will be done on C.S. Peirce's *Collected Papers (Peirce, 1931)*, and both the construction of the corpus and the specifics of applying CoFiH on it will be presented in section §5. After presenting the results (section §6), we discuss how it reflects on our specific hypothesis in section §7.

2. Previous work

Existing approaches of concept modeling in computer science fall into two broad camps. On one hand, the keyword-based approaches identify concepts to keywords or keyword sets. As it simplifies treatment, it is implicitly used in many works, among which those based on the word-space model (Salton & McGill, 1983) or on word co-occurrence (Lund & Burgess, 1996). On the other hand are methods which see text segments or words as expressions of meaning dimensions which often take the name of “concept” or “topic”, and which are often refered to as “topic

models". Among them are latent models (LSA, LDA, etc.) and explicit semantic analysis or ESA (Gabrilovich & Markovitch, 2007).

Keyword-based approaches rely on words having the same meaning no matter what the context is, and as such, are poor models for concepts when context changes enough for words to take on different meanings. Latent topic models, on the other hand, model objects (topics) which are much broader than concepts, and are typically associated with a plurality of related concepts. As such, they model a different object. ESA does claim to model concepts themselves through their Wikipedia articles; however, recent works (Anderka & Stein, 2009; Gottron, Anderka, & Stein, 2011) seem to show that this claim doesn't explain ESA's performances, and equates ESA to a LSA variant (Liu & Wang, 2012).

3. Hypotheses

Our approach to the problem of finding relevant textual segments for conceptual analysis relies on a set of hypotheses, which guide the the heuristic we developed (cf. §4).

Hypothesis 1. *Is relevant to a concept any textual unit that participates to a topic where the said concept is expressed.*

Concepts can be involved in various discussions and discursive contexts, and thus be associated with very different ideas. The stranger plays a very different role whether it is a hero figure in bildungsroman or a threatening Other in news items, but it remains the same concept. We shall say that these two discursive contexts are two topics of the concept STRANGER. Topics regroup propositions or textual fragments which are about the same thing, as may be expressed by e.g. a proposition (Van Dijk, 1977, pp. 131–42).

Furthermore, concepts are involved in a way or another in all of the expression of the topics in which they participate, as the concepts involved in a topic collectively constitute it. Whether we consider that topics' structure is propositional, like Van Dijk, or otherwise, topics put concepts in relationships with other concepts, and these relationships are the ones potentially relevant to conceptual analysis.

Hypothesis 2. *The topic is a latent variable of the text.*

In other words, the probability of apparition of a lexeme in a segment is a function of the presence or absence of the various topics of the corpus.

A text is made of lexemes—neither theme nor concept can directly be apprehended. To "read" them from textual data, we suppose that there is an undisclosed relationship between latent variables that structure the text (such as topics) and the text's observable features. As the former are presumed to condition the latter, it can be deduced that we can estimate from lexemes the probability that a given topic is latently present.

Data: D, q, α
Result: R
1 $l \leftarrow \alpha N$;
2 $E \leftarrow \{d_i | q \cdot d > 0, d_i \in D\}$;
3 $C \leftarrow \text{Partition}(E)$;
4 **foreach** $c_k \in C$ **do**
5 $T_k \leftarrow \{t_i | \text{tScore}(D, c_k) \text{ in the top } l\}$;
6 $D' \leftarrow (\{b_i | b_i \in D^T, i \in T_k\})^T$;
7 $D'_k \leftarrow \{d'_i | d'_i \in D', i \in c_k\}$;
8 $\mu_k \leftarrow \frac{1}{c} \sum_{d' \in D'} \text{Distance}(d'_i, \bar{d}'_k)$;
9 $\sigma_k \leftarrow \sqrt{\frac{1}{c_k} \sum_{d'_i \in D'_k} \|\text{Distance}(d'_i, \bar{d}'_k) - \mu_k\|^2}$;
10 $J_k \leftarrow \{d_i | \text{Distance}(d_i, \bar{d}') \leq (\mu_k + 2\sigma_k), d_i \in D\}$;
11 $R \leftarrow \cap_k J_k$

Algorithm 1: CoFiH

While we do not make any supposition as the nature of this relationship, we suppose that our corpus is relatively uniform in terms of pragmatics (agents involved, language games, etc.). As such, topics ought to be the main structuring latent variable. In this, we follow the tradition of topic models (e.g. Blei et al., 2003; Landauer, Foltz, & Laham, 1998)

Hypothesis 3. *The lexical footprints of segments expressing a particular topic follow a normal distribution with, as its mean, a prototypical vector which represents the said topic.*

This hypothesis compounds three related assumptions. One is that segments which are lexically similar are likely to express the same topic(s). As topics are made of recurrent concepts and motifs, it is common sense that they would also feature recurrent content vocabulary. The other is that topics' expression on the vector space of word expressions will translate into clusters of textual segments: as those segments express the same topic with some variations, they will appear close to each other. And the last one is that the factors which differentiate those segments and yield the variation among them can be modeled as gaussian noise.

Hypothesis 4. *The segments expressing the canonical expression of a concept (the word(s) with which designates them most frequently) is representative in terms of topics expressions.*

While a concept's expression is likely to change depending on the context, we may assume that its canonical expression is likely to recur if only because authors might want to activate the same bodies of knowledge.

4. Method

In this section, we present CoFiH (Concept-Finding Heuristic), an algorithm which exploits the aforementioned hypotheses in order to retrieve segments that can be of use to a conceptual analysis.

Topics are first extracted from the sub-corpus of segments that contain the canonical expression of the concept being queried (restricting ourselves to this subcorpus makes for faster calculations and more accurate topic modeling). For each topic, a distribution of the textual segments expressing it is modeled according to hypothesis 3, and likely candidate segments are extracted. The result, the union of all recalled segments, is the set of candidates for conceptual analysis.

This method takes the form of a data processing sequence that implements our algorithm, CoFiH (cf. Algorithm 1). Firstly, the corpus is converted into a word space model (Salton & McGill, 1983). After pretreatment (which includes exclusion of non-alphabetical characters, word tokenization, stopwords filtering and lemmatization), the text is segmented and segments are represented as vectors $d_i = \{a_{ij}, \ldots, a_{iN}\} \in D$ where a_{ij} is the word count for the j^{th} word in the i^{th} segment and N is the number of distinct words. The queried concept Q is expressed as a canonical expression which can be translated into a set of words: e.g. the concept BODY can be translated as {"body", "bodily", …} and represented as a vector $q = \{a_{q_1}, \ldots, a_{q_N}\}$ on the vector space constructed by D. It can be used to extract the set $E \subset D$ of segments which contain at least one of the words from Q (Algorithm 1, line 2).

An unsupervised clustering technique is then used to partition E (line 3), and for each cluster c_k, *t*-scores for each word are calculated. A matrix D'_k is then constructed with the vectors for segments in c_k whose *t*-scores (cf. Manning & Schütze, 1999) are with the top l (lines 6 and 7). A vector $\bar{d}'_k$, mean of all vectors $d'_i \in D'_k$, is constructed as a prototype of class c_k, and the mean and standard deviation of the cosine distance between vectors $d'_i \in D'_k$ and $\bar{d}'_k$ are calculated (lines 8 and 9). Cosine distances between vectors $d_i \in D$ and $\bar{d}'_k$ are then calculated on the attributes D and D'_k have in common. Set J_k is then constructed with vectors $d_i \in D$ which are within a distance $\mu_k + 2\sigma_k$ of $\bar{d}'_k$, and it is called the extension of topic k. The end result, i.e. the set of segments pertinent to a conceptual analysis of concept Q, is the union of all J_k, with the possible exception of those topics which the expert will have deemed unrelevant to her analysis.

5. Experimentation

The chosen corpus for our experimentation was the *Collected papers* by C. S. Peirce, which was studied in a previous work (J.-G. Meunier & Forest, 2008). It consists in 4,965 sections in 8 volumes. For the needs ot this experiment, it was tokenized into sentences and words using NLTK, stopwords (from NLTK's english stopwords list) were removed, and remaining words were lemmatized using the WordNet lemmatizer. Segments were then made by splitting sections

	Topic 0	Topic 1	Topic 2	Topic 3	Topic 5	Topic 6
N	*1751*	*4*	*2081*	*107*	*668*	*2994*
Words with highest t-score	one	evade	would	number	law	may
	law	metaphysically	one	zero	nature	one
	upon	subsumption	law	precept	phenomenon	upon
	must	inherence	upon	feeling	general	two
	may	dualistic	say	average	one	every
	would	ruling	could	rectangle	physical	law
	every	impose	must	wide	mind	must
	two	defender	might	hundred	upon	general
	thing	imposed	fact	doublet	force	say
	first	literal	two	teeth	motion	fact
	general	governs	thing	eight	fact	first
	fact	affirmation	mind	region	first	proposition
	another	rejecting	true	atomicules	must	certain
	idea	arduous	certain	experiment	time	world

Table 1: Most associated words for each non-empty topic of LAW

in smaller groups of 3 consecutive sentences[1]. Furthermore, segments with less than 30 words or no content word were removed, leaving us with a total of 14,490 segments. A test subset of 98 segments from 40 different sections was tagged by 4 experts, who were asked to identify and write down 5 concept for each segment. Input tags were lemmatized using WordNet and stored in a database.

Our data processing sequence was applied with, as query, 285 different concepts which experts employed to tag at least one segment and corresponded to a content word in the corpus. While most tags were associated by experts to only one segment, some were used more frequently, with "logic" being the most frequent with mentions on 20 different segments.

For the clustering method, we chose to use the *k*-means algorithm, initialized with kmeans++. For each concept query, the number of clusters was determined using the $f(K)$ method (Pham, Dimov, & Nguyen, 2005) with the cosine distance.

Our tagging method guarantees that, in the mind of the tagger, a given segment is relevant to a concept, but absence of tag doesn't indicate its irrelevance. As such, recall (the proportion of segments known to be relevant that are retrieved by the algorithm) can be measured. It will be measured for both CoFiH and our baseline, a heuristic which consists in retrieving segments which contain the canonical expression of the query concept.

Precision, however, cannot be measured, as it requires that we know all of the segments relevant to a given concept in the test subset[2]. This threatens to make recall useless, as high recall can be

[1]A portion of this corpus was annotated for purposes of validation, and large segments would have demanded too much work from the annotators. As such, smaller segments were deemed necessary.

[2]For the purpose of measuring precision, a new round of tagging will be done by the end of spring 2016, making these results available in time for presentation in June.

"gamed" by recalling more segments, even randomly, in hopes of stumbling upon more of the tagged segments. Thus, to evaluate the quality of our subset, the concept of LAW[3] was analyzed based on topics' most discriminative words (as indicated by the *t*-score) and segments randomly drawn from them (10 segments per topic).

6. Results

CoFiH retrieved an average of 3,230 segments against 255 for the baseline. The average size of the intersection of those two sets of segments was 254, indicating that CoFiH almost always retrieves what has been retrieved by the baseline.

6.1. Recall

CoFiH's average recall measures as 69,3%, whereas the baseline's is at 56,9%, when averaged along concepts. Weighting concepts proportionally to the amount of tagged segments does not make an important difference: CoFiH's recall drops to 67,3% while baseline recall drop to 52,1%.

6.2. Analysis of LAW

CoFiH finds 7 topics for the query "law", one of which retrieves no segment, and cannot therefore be analyzed. For each topic, the most associated words as measured by the *t*-score are shown in table 1.

Themes can be inferred from words with high *t*-scores.

The bigger classes (0, 2 and 6), which all contain a majority of the retrieved segments, are all characterized by words which look like function words, but can actually expressed modality ("would", "might", "may", etc. for possibility; "every", "must", "upon", etc. for necessity). While these words are often merely functional, they often play a role as the author is illustrating a law or deducing an hypothesis (which, as Peirce sees it, is done by making use of the laws of thought for the purpose of discovering laws of categorization).

Topic 0 and 6 seem to distinguish themselves as some of the vocabulary deals with generality ("general", "idea", "proposition"). A recurrent thread in the *Collected Papers* is the production of general thought entities from individual ones, as, for example, the ones that are presented to us by our senses. This process, for Peirce, depends on laws both for discovering the generality and for formulating it (what unites individuals is a law). For instance, in topic 0, we find:

> "If on the other hand, we find that as soon as the form is prevented from manifestation in one shape it immediately reappears in another shape, and especially if it shows a power of spreading and of reproducing itself, these phenomena may be considered as evidence of genuine vitality and fundamental reality in the form of the law." (*CP*, §7.469)

Topic 2, on the other hand, is further characterized by words indicating themes from epistemology ("true", "certain") and psychology ("mind"), which might be explained by the high

[3]The choice of this concept is arbitrary.

prevalence of passages discussion the problem of universals in medieval philosophy, such as this one:

> To say that the conceptualistic and nominalistic theories are both true at once, is mere ignorance, because their numerical results conflict. A conceptualist might hesitate, perhaps, to say that the probability of a proposition of which he knows absolutely nothing is 1/2, although this would be, in one sense, justifiable for the nominalist, inasmuch as one half of all possible propositions (being contradictions of the other half) are true; but he does not hesitate to assume events to be equally probable when he does not know anything about their probabilities, and this is for the nominalist an utterly unwarrantable procedure. A probability is a statistical fact, and cannot be assumed arbitrarily. (*CP*, §8.4)

Among the smaller topics, topic 1 is about dualism between the laws of physics and the laws of thought. As for topic 3, as terms like "number", "average" and "rectangle" indicate, it is concerned with laws in mathematics and geometry:

> The results of experiments would all be expressed in two sets [of] numbers, one showing the percentage of errors and the other the feeling of confidence, in the attempts at discrimination under different circumstances. Those numbers were then subjected to mathematical discussion, according to the established principles of such work; and from them a law was deduced. (*CP*, §7.546)

Topic 5, finally, is characterized by its concern for laws of physics, as witnessed by terms such as "nature", "physical", "force" and passages such as this one:

> That reason is that the laws of motion make velocity of rotation to be something absolute and not merely relative. Now velocity is the ratio of the amount of a space-displacement to the amount of time in which this displacement takes place. (*CP*, §7.486)

In summary, in the *Collected Papers*, the concept of LAW plays a fundamental role in inference and conceptuality, as it enables both discovery and formulation of generality. It is expressed as much in the domain of physics as the domain of thought (despite the dualistic divide) and, correspondingly, in geometry and arithmetics (hence the importance of mathematics).

Of the 54 segments drawn (only four could be drawn from topic 1), only 1 was judged irrelevant, and 49 were judged relevant to the concept of LAW.

7. Discussion

In terms of recall, CoFiH solidly outperforms the baseline, scoring 69% compared to 57%. Such improvement seems important; however, as CoFiH retrieves more than 10 times the amount of segments that the baseline recalls, one might wonder if it isn't simply due to chance.

This is where our analysis may help us. While reading excerpts from each topic generated by CoFiH for LAW, with only 1 (>2%) being judged irrelevant, we judged that at least 90% of the segments seemed relevant to the analysis. Those segments went on to facilitate an *ad hoc* interpretation of this concept in Peirce's works.

However, interpretative results are anecdotal: what goes on with one concept isn't necessarily indicative of what goes on with the others. Perhaps more importantly, natural language interpretation is liable to confirmation bias (Wason, 1960): we may be reading too much into what is given to us, and abusively categorize segments as relevant even though they really aren't. On the one hand, some annotations might provide data about false positives which could go a long to address those concerns. However, on the other hand, a proper test of a technique to assist textual analysis can only be in the context of interpretation, and as such, it might not be possible to completely put those worries to rest.

Another limit rests with the simplifications that lie in our hypotheses. The topic model that hypothesis 3 describes is a coarse simplification of modern state-of-the-art models (Blei et al., 2003). Hypothesis 4 seems in tension with hypothesis 1 in that we might expect a concept to express itself with different lexemes in different topics. The force of these hypotheses is that they enable the conception of very simple algorithm. However, the success of the latter suggests that the former could be refined, allowing for a more powerful model in future work.

8. Conclusion

Our method illustrates how the notion of topic can be exploited to retrieve textual contexts which are likely to be relevant for a conceptual analysis. As a proof of concept, we have shown how the results represented in the form of strongly associated words (table 1) and randomly drawn segments can be used to make interpretations of the topics related to a concept.

Both the high recall and obvious potential of the data it produces for conceptual analysis suggest that it shows promise. However, further validation might be necessary to evaluate the quality of the retrieved subset of segments, particularly in terms of precision.

References

Anderka, M., & Stein, B. (2009). The ESA Retrieval Model Revisited. In *Proceedings of the 32Nd International ACM SIGIR Conference on Research and Development in Information Retrieval*. ACM, New York, USA, pages 670-1.

Blei, D. M., Ng, A. Y., & Jordan, M. I. (2003). Latent dirichlet allocation. *The Journal of Machine Learning Research*, 3: 993-1022.

Chartier, J.-F., & Meunier, J.-G. (2011). Text mining methods for social representation studies. *Papers on Social Representation*, 20: 37.1-37.47.

Chartier, J. F., Meunier, J. G., Danis, J., & Jendoubi, M. (2008). Le travail conceptuel collectif: une analyse assistée par ordinateur du concept d'ACCOMMODEMENT RAISONNABLE dans les journaux québécois. In S. Heiden & B. Pincemin, editors, *JADT 2008 : actes des 9es Journées internationales d'Analyse statistique des Données Textuelles, Lyon.* Presses universitaires de Lyon, pages 297-307.

Danis, J., Meunier, J. G., Chartier, J. F., Alrahabi, M., & Desclés, J. P. (2010). Classification automatique et stratégie d'annotation appliquées à un concept philosophique: la dimension psychologique du

concept de LANGAGE dans l'œuvre de Bergson. In S. Bolasco, I. Chiari, & L. Giuliano, editors, *JADT 2010: 10th International Conference on Statistical Analysis of Textual Data*, pages 49–60.

Deerwester, S. C., Dumais, S. T., Landauer, T. K., Furnas, G. W., & Harshman, R. A. (1990). Indexing by latent semantic analysis. *JASIS*, *41*(6): 391–407.

Estève, R. (2008). Une analyse quantitative de la morale chez Vladimir Jankélévitch. In S. Heiden & B. Pincemin, editors, *JADT 2008 : actes des 9es Journées internationales d'Analyse statistique des Données Textuelles, Lyon*.

Fayyad, U., Piatetsky-Shapiro, G., & Smyth, P. (1996). From data mining to knowledge discovery in databases. *AI Magazine*, *17*(3): 37.

Gabrilovich, E., & Markovitch, S. (2007). Computing Semantic Relatedness Using Wikipedia-based Explicit Semantic Analysis. In M. M. Veloso, editor, *IJCAI 2007, Proceedings of the 20th International Joint Conference on Artificial Intelligence, Hyderabad, India, January 6-12, 2007*, pages 1606-11.

Gottron, T., Anderka, M., & Stein, B. (2011). Insights into explicit semantic analysis. In C. Macdonald, I. Ounis, & I. Ruthven, editors, *Proceedings of the 20th ACM Conference on Information and Knowledge Management, CIKM 2011, Glasgow, United Kingdom, October 24-28, 2011*, pages 1961-4

Haslanger, S. (2012). *Resisting Reality: Social Construction and Social Critique*. Oxford University Press.

Jackson, F. (1998). *From Metaphysics to Ethics: A Defence of Conceptual Analysis*. Oxford University Press.

Landauer, T. K., Foltz, P. W., & Laham, D. (1998). An introduction to latent semantic analysis. *Discourse Processes*, 25: 259–284.

Laurence, S., & Margolis, E. (2003). Concepts and Conceptual Analysis. *Philosophy and Phenomenological Research*, 67(2): 253–282.

Lebart, L., & Salem, A. (1994). *Statistique textuelle*. Dunod.

Liu, C., & Wang, Y.-M. (2012). On the Connections Between Explicit Semantic Analysis and Latent Semantic Analysis. In *Proceedings of the 21st ACM International Conference on Information and Knowledge Management*, pages 1804-8. ACM, New York, USA.

Lund, K., & Burgess, C. (1996). Producing high-dimensional semantic spaces from lexical co-occurrence. *Behavior Research Methods, Instruments, & Computers*, 28(2): 203-8.

Manning, C. D., & Schütze, H. (1999). *Foundations of Statistical Natural Language Processing*. MIT Press.

Meunier, J.-G., & Forest, D. (2008). Computer assisted conceptual analysis of text: the concept of mind in the Collected papers of C.S. Peirce. In *Actes du colloque international Digital Humanities*. Oulu, Finlande.

Meunier, J. G., Forest, D., & Biskri, I. (2005). Classification and Categorization in computer-assisted reading and text analysis. In H. Cohen & C. Lefebvre (Eds.), *Handbook of Categorization in Cognitive Science*. The Hague: Elsevier, pages 955-78.

Peirce, C. S. (1931). *Collected Papers of Charles Sanders Peirce*. Cambridge, Harvard University Press.

Pham, D. T., Dimov, S. S., & Nguyen, C. D. (2005). Selection of K in K-means clustering. *Proceedings of the Institution of Mechanical Engineers, Part C: Journal of Mechanical Engineering Science*, 219(1): 103–119.

Sainte-Marie, M. B., Meunier, J.-G., Payette, N., & Chartier, J.-F. (2011). The concept of evolution in the Origin of Species: a computer-assisted analysis. *Literary and Linguistic Computing*, 26(3): 329–334.

Salton, G., & McGill, M. J. (1983). *Introduction to modern information retrieval*. McGraw-Hill.

Van Dijk, T. A. (1977). *Text and Context: Explorations in the Semantics and Pragmatics of Discourse*. Addison-Wesley Longman Limited.

Wason, P. C. (1960). On the failure to eliminate hypotheses in a conceptual task. *Quarterly Journal of Experimental Psychology*, 12(3): 129–140.